Conception d'un atelier de génie documentaire pour les banques de données juridiques

Adel Saadoun, Claude Belair, Jean-Louis Ermine, Jean-Marc Pouyot

---

Version française non publiée

Article paru en version anglaise
"A knowledge engineering framework for intelligent retrieval of legal case studies"
dans
Artificial Intelligence and Law, 1-27, 1997, Kluwer Academic Publishers

Conception d'un atelier de génie documentaire pour les banques de données juridiques

Adel Saadoun[**/***], **Claude Belair**[*], **Jean-Louis Ermine**[**], **Jean-Marc Pouyot**[***],

[*]*JURIS DATA,*
123 rue d'Alésia 75678 Paris
[**]*Commissariat à l'Énergie Atomique*
DIST/SMTI
Groupe Gestion des Connaissances
Centre d'Études de Saclay
91191 Gif sur Yvette Cedex
[***]*Scalaire, rue Lafaurie Montbadon, 33000 Bordeaux*

RESUME : Juris-Data est une des plus importantes bases documentaires juridiques de France. Les documents sont indexés par une classification juridique élaborée par Juris-Data. Des techniques de génie cognitif ont été utilisées pour réaliser une interface intelligente pour la recherche de cas, basée sur cette classification. Le but de ce système est d'aider l' utilisateur à trouver le cas juridique le plus proche de celui qui le concerne. Cette approche s'est révélée fructueuse, mais pour la généraliser à d'autres bases, il est nécessaire d'extraire une classification juridique des documents de la base. Pour cela, une méthodologie de construction de telles classifications a été élaborée, en même temps qu'une méthodologie de construction d'index. Le projet dans son ensemble a abouti à la réalisation d'un "atelier de génie documentaire juridique" basé sur les expériences acquises et les méthodologies développées, qui est un ensemble d'outils informatisés qui supporte tout le cycle de vie du document juridique, de son traitement par un analyste juriste, jusqu'à sa consultation par un client.

MOTS-CLES : Banques de données juridiques, recherche d'informations, intelligence artificielle, génie cognitif, bases documentaires

ABSTRACT : Juris-Data is one of the most important legal document base in France. The documents are indexed by a legal classification elaborated by Juris-Data. Knowledge engineering has been used to design an intelligent interface for case retrieval, based on that classification. The aim of that system is to help users to find the legal case relevant to their own case. The approach has been successful, but to generalize it to other bases, it needs to extract a legal classification from the documents of the base. Thus, a methodology for designing such classification has been elaborated, together with a methodology for index construction. The whole project led to the implementation of an "legal document engineering workbench", based on the acquired experiences and methodologies, which is a set of computerized tools supporting the whole life cycle of legal documents, from the processing by lawyers/analysts to the consultation by clients.

KEY WORDS: Law Database, Information Retrieval, Artificial Intelligence, Knowledge Engineering, Document Base

1. Introduction

Dans le domaine juridique, la surabondance des informations fournies par une documentation qui ne cesse de croître rend difficile le suivi et la gestion manuelle des documents. L'accroissement, la diversité et la richesse des documents à gérer ont obligé les documentalistes à recourir à un traitement informatique. On a ainsi assisté à un transfert de la documentation du support papier sur support magnétique. Les besoins des utilisateurs, qui sont ainsi passés de l'identification d'un document à l'identification d'une information contenue dans un document électronique, ont conduit au développement et à la diversification des instruments de recherche d'informations. De cette approche nouvelle de la documentation sont nés les premiers Systèmes de Recherche d'Informations (SRI).

A ce jour, la plupart des grandes bases documentaires ont été souvent conçues du seul point de vue quantitatif de la gestion des documents. On a donc surtout assisté à l'automatisation de la documentation existante, ou à la constitution de documentation nouvelle selon les méthodes déjà existantes pour la documentation papier.

Le domaine juridique comme tout autre domaine, séduit par l'idée de pouvoir gérer automatiquement sa documentation, se lance à son tour dans la conception de ses propres bases documentaires. Les premiers développements, limités à une simple transposition des logiciels et des réflexions existant pour d'autres types de documents (techniques, scientifiques, bibliographiques, etc.), se sont avérés inadaptés pour les documents des juristes.

En effet, plus le domaine des connaissances est riche (ce qui est le cas pour le domaine juridique) plus les besoins des utilisateurs sont ponctuels et précis. Les instruments documentaires classiques ainsi que les bases documentaires conçues du seul point de vue quantitatif de la gestion des documents, ne satisfont qu'assez mal ce genre de besoin.

Face à cette inadéquation des techniques et approches documentaires traditionnelles, il importe de penser à une autre conception de la documentation juridique. Cette conception se doit d'organiser la documentation, d'analyser son contenu et de construire des instruments documentaires selon une méthodologie purement cognitive, c'est à dire proche non pas de la machine, mais des modes de compréhensions cognitifs des utilisateurs. La compétence des documentalistes étant restreinte au domaine documentaire, il s'agit d'un travail théorique qui nécessite la collaboration de juristes, d'informaticiens et de spécialistes d'Intelligence Artificielle [Bou91].

Le but du travail présenté ici est de concevoir un Atelier de Génie Documentaire Juridique (AGDJ), c'est à dire un ensemble d'outils informatisés, fédérés par une méthode d'utilisation structurée et destinés à aider, à uniformiser et rendre cohérente toute la chaîne de production et d'utilisation de documents juridiques, de l'analyse et de la saisie des documents juridiques jusqu'à l'interface de recherche d'informations.

Dans cet article, nous nous intéressons à l'une des composantes de l'AGDJ, à savoir la construction d'interfaces "intelligentes" pour l'aide à la recherche d'informations dans le domaine juridique. De telles interfaces, s'appuyant sur de la connaissance propre aux documents d'une base documentaire juridique donnée, doivent guider l'usager dans ses recherches et l'assister dans sa réflexion.

Pour la réalisation de ce travail, nous avons procédé par étapes. L'approche envisagée se compose de deux parties :

A) La construction de l'interface d'interrogation pour une base documentaire juridique structurée appelée JURIS-DATA (JD). On entend par structurée, toute base documentaire disposant au moins d'une structure classificatoire assez pertinente et représentative de ses documents. La démarche entreprise pour la réalisation de cette première partie comprend plusieurs phases :

A1) Une analyse fonctionnelle de l'activité du groupe JD (faite avec le langage SADT). Cette analyse a permis d'identifier les tâches susceptibles d'être améliorées dans la chaîne documentaire, notamment par l'utilisation de systèmes à base de connaissances. Ces tâches sont : l'aide à l'analyse des documents juridiques et l'aide à la recherche d'informations

A2) La construction d'une base de connaissances propre à JD. Il s'agit de spécifier les connaissances, juridiques, documentaires et heuristiques contenues dans les documents (en ligne) et impliquées dans leur création (Connaissance statique).

A3) La spécification de la stratégie d'un expert quand il résout une tâche de recherche documentaire. Cette phase menée en collaboration avec un expert de la base (Mr C. Belair) et les futurs utilisateurs précise, entre autres, la manière selon laquelle la base de connaissances sera utilisée afin de réaliser une recherche documentaire (Connaissance dynamique).

A4) La conception de l'interface intelligente d'aide à la recherche d'informations en tenant compte de certaines contraintes matérielles (la base est essentiellement consultée par les professionnels sur le réseau français du Minitel) et ergonomiques (certains utilisateurs occasionnels ne sont pas accoutumés aux interfaces sophistiquées).

A5) L'implantation, le test et la validation de l'interface avant le développement opérationnel.

B) La généralisation de l'interface d'interrogation pour d'autres bases documentaires juridiques moins structurées, où on ne dispose pas de structures de types JD.

Conformément à la définition du label "structurée" énoncée auparavant, la différence entre deux bases documentaires, l'une structurée et l'autre moins, réside dans l'existence ou pas d'une structure classificatoire des documents. C'est cette classification, qui dans notre approche, constitue l'ossature de la base de connaissances sur laquelle repose l'interface d'interrogation.

Compte tenu de cette hypothèse, une méthodologie de construction de structures documentaires (structures classificatoires et index alphabétique) a été élaborée et testée. Partant de l'ensemble des documents de la base, un expert juridique procède selon une méthode organisée soit à la construction de la classification soit à la construction de l'index. Grossièrement, construire une classification consiste en trois étapes :

B1) Extraction de la connaissance: Cette étape permet d'extraire dans les documents, les unités lexicales qui traduisent des concepts juridiques pertinents. Ces concepts, qu'ils soient implicites ou explicites, représentent la connaissance juridique qu'intègre la base documentaire.

B2) Normalisation de la connaissance extraite: il s'agit d'unifier, de standardiser le vocabulaire exprimant la connaissance. Cette standardisation repose sur un choix du vocabulaire dans un corpus constitué, des unités lexicales extraites lors de l'étape précédente et des synonymes, analogues, allotaxies et antonymes correspondants.

B3) Organisation de la connaissance normalisée en structures classificatoires juridiques: Au cours de cette étape, l'expert procède à l'établissement de liens hiérarchiques et autres entre les différents concepts normalisés. Ces concepts reliés entre eux et structurés en réseaux sémantiques représentent la structure classificatoire des documents de la base.

Comme on vient de le voir, cette approche réclame l'intervention d'un expert juriste. Dans le cas où aucun expert n'est disponible, diverses techniques automatiques d'extraction et de structuration de connaissances à partir de documents sont mises en oeuvre. Ces techniques proviennent de plusieurs domaines tels que la linguistique, l'Intelligence Artificielle, le documentaire, etc.

2. JURIS-DATA : Une base documentaire juridique structurée

2.1. Fonctionnement de J-D

JURIS-DATA constitue l'une des plus anciennes et des plus importantes bases documentaires juridiques françaises. Cette base, fondée sur la méthode de l'analyse et abstract et consultable sur le réseau français du Minitel, réunit des centaines de milliers de documents recueillis dans plusieurs sources du droit. On retrouve de la jurisprudence[1], de la doctrine[2] et des réponses ministérielles[3]. Face à l'accroissement de plus en plus important du nombre de documents juridiques produits chaque année en France, il est quasiment impossible de prévoir un système documentaire efficace sans une sélection préalable de l'information. Pour JD, la sélection de cette documentation (notamment la jurisprudence) se doit de respecter deux aspects :

- Un aspect qualitatif : il s'agit d'éliminer toutes les décisions de justices qui, au regard de la matière évoquée, sont d'un intérêt juridique restreint.

- Un aspect quantitatif : il s'agit d'éliminer des arrêts pertinents mais répétitifs dont l'amoncellement dans la base, loin de favoriser l'émergence de l'information, contribue, au contraire, à l'étouffer.

Afin de mieux satisfaire ses utilisateurs (spécialistes et généralistes), JD a décidé de dépasser le niveau strictement documentaire des banques de données traditionnelles. L'information brute (telle que proposée par la méthode du texte intégral) ne suffit plus à l'utilisateur lequel attend de la base qu'elle le guide dans ses recherches, qu'elle l'assiste dans sa réflexion. Par conséquent, l'un des objectif prioritaires de JD est de mettre à la disposition de l'utilisateur non pas des documents en texte intégral (dits primaires) mais plutôt des documents de synthèse appelés "abstracts". Ces abstracts étant destinés à des juristes de formation classique, il importait de reprendre les cheminements logiques juridiques habituels. Cependant, pour que cette documentation soit parfaitement compréhensible, complète et diffusable, l'information qu'elle véhicule doit être non seulement structurée, canalisée, homogène, mais encore analysée.

[1] C'est l'ensemble des décisions de justice rendues par les différentes juridictions
[2] C'est l'ensemble des opinions des juristes sur des sujets n'ayant pas d'existence légale
[3] C'est l'ensemble des réponses des ministres du gouvernement aux questions posées par les représentants du peuple

Une fois sélectionnées, ces documents primaires sont donc analysées suivant des structures préétablies appelées Structures d'Analyse (SA). Une SA est une décomposition du droit en différentes étapes allant du droit au fait. C'est un découpage de la science juridique inspiré à la fois, de la classification légale établie par le législateur dans les différents codes (civil, du travail, pénal, etc.), et de la doctrine qui traite des sujets n'ayant pas d'existence légale (droit des affaires, droit de l'environnement, etc.). Celle-ci est organisée à la manière des classifications naturelles décrites dans [Vog 89]. Il s'agit d'une classification à cinq niveaux où le premier niveau rend compte de la matière juridique (JD en compte une trentaine: Assurance, Bail commercial, Construction, etc.), les autres sont des spécialisations progressives du niveau hiérarchiquement supérieur. Un exemple de SA est donné par la figure 2.1.

Matière juridique: ASSURANCES
Niveau générique: ASSURANCE EN GENERAL, OBLIGATION DE L'ASSUREUR,
Niveau spécifique:
-- ARTICLE L.113-1 DU CODE DES ASSURANCES, EXCLUSION DE GARANTIE,
* EXCLUSION FORMELLE ET LIMITEE (OUI/NON),.....
* CARACTERE GENERAL DE L'EXCLUSION (OUI/NON),.....
* >>DEFINITION/DISTINCTION<< DE L'EXCLUSION,.....
--ARTICLE L.113-1 ALINEA 2 DU CODE DES ASSURANCES, EXCLUSION DE GARANTIE,
* FAUTE >>INTENTIONNELLE/DOLOSIVE<< DE L'ASSURE (OUI/NON),.....
--ARTICLE L.113-5 DU CODE DES ASSURANCES,
* EXECUTION DU CONTRAT DANS LE DELAI PREVU (OUI/NON),.....
--POINT DE DEPART DE LA RESILIATION >>DU CONTRAT/DE LA POLICE<<,.....
--REFUS DE RENOUVELLEMENT>>DE LA POLICE/DU CONTRAT<< PAR L'ASSUREUR,.....
>>RESPECT/NON RESPECT<< DE SES OBLIGATION (OUI/NON),.....

Figure 2.1 : Exemple de Structures d'analyse pour JD

Une SA indique pour chaque type de problème juridique (ou matière juridique), le vocabulaire à utiliser, le séquencement ainsi que la nature des informations à extraire d'un document juridique primaire. C'est en s'appuyant sur ces SA que les analystes parviennent à rédiger les abstracts. Un abstract est donc défini comme étant le résultat de l'analyse d'un document primaire. Il reprend d'une manière précise, concise et organisée, les principaux concepts juridiques implicites et/ou explicites qu'évoque le document à analyser.

Après la phase d'indexation et d'intégration à JD, les abstracts sont enfin prêts à la diffusion. L'interrogation de la base s'effectue en langage naturel grâce au SRI SYSDEX (développé par la société SCALAIRE).

2.2. Démarche suivie

Comme on vient de le voir, JD n'est pas seulement un ensemble de documents électroniques, mais un système complexe, évolutif, aux aspects multiples et dont la maîtrise ne tient pas uniquement à une ou plusieurs techniques, mais à une méthode organisée. Dès lors et pour mieux appréhender le problème, notre démarche a consisté en un véritable travail de gestion des connaissances. Il s'agit d'expliciter, pérenniser et transmettre le patrimoine de connaissances qui s'est accumulé tout au long de la vie du groupe JD. Les modèles obtenus permettent d'approcher de manière pertinente les problèmes de gestion de connaissances au sein du groupe, et d'éventuellement proposer des solutions [Bru 94]. Les différentes étapes de notre démarche sont :

- La modélisation de l'activité
- L'identification des points critiques dans l'activité
- La proposition de solutions
- La modélisation cognitive

2.2.1. Modélisation de la connaissance

Notre première tâche a été un travail de mise en contexte de l'activité, afin d'appréhender correctement la complexité du problème de la documentation dans sa globalité. Nous avons donc réalisé une analyse fonctionnelle détaillée et complète de toute la chaîne de production et d'exploitation des données JD. Cette vue éclatée de la chaîne faite avec le langage graphique S.A.D.T [IGL 89], nous a révélé l'existence d'une réelle connaissance ou expertise, complexe et structurée traduisant un savoir faire certain, propre à l'unité JD. Mais elle a aussi dévoilé les maillons où une intervention pouvait améliorer les potentialités du système déjà en place.

2.2.2. Identification des points critiques

Actuellement, la tâche de rédaction d'abstracts est complètement manuelle. Bien que les analystes doivent se conformer strictement aux consignes de résumé apportées par les SA, cela n'empêche pas certaines formes d'omissions, de subjectivité, d'anticipation de certains points, et ainsi de fausser une partie de l'analyse. Or, quand on sait que la cohérence, l'homogénéité et l'efficacité de JD dépendent essentiellement de la qualité de l'analyse des documents primaires, autrement dit des abstracts, on ressent la nécessité de fournir aux analystes des outils adaptés afin de minimiser tout risque d'erreur.

Si la technique de l'analyse et abstract adoptée par JD, comparée à celle du texte intégral, a le mérite d'unifier le vocabulaire, elle présente un inconvénient pour l'utilisateur non spécialiste ne connaissant pas les termes choisis pour la structuration de l'abstract. En effet, face à un écran vide, il est indispensable que l'utilisateur ait déjà bien analysé son problème et qu'il ait acquis le langage qui doit servir à l'interrogation. Pour pallier à cette éventuelle absence de préparation préalable chez certains utilisateurs, notamment les néophytes, il importe de concevoir des interfaces "intelligentes" d'aide à la recherche d'informations. Celles-ci se doivent de guider l'utilisateur dans ses recherches et de l'assister dans sa réflexion.

Compte tenu des deux réflexions précédentes, les points d'interventions déterminés sont :

- Un outil d'aide à la rédaction d'abstracts
- La conception d'interfaces d'interrogation "intelligentes" et plus adaptées au raisonnement juridique, pour l'aide à la recherche d'informations. C'est seulement ce dernier point qui sera détaillé dans la suite de cet article.

2.2.3. Proposition de solutions

A partir du moment où on a vu que les systèmes envisageables étaient nécessairement basés sur la connaissance, l'approche Génie Cognitif peut être utilisée. C'est une approche méthodologique, permettant d'analyser un ensemble de connaissances propres à un ou plusieurs experts au sein d'un groupe afin d'en obtenir un modèle structuré, cohérent et opératoire; ce travail pouvant se concrétiser par un logiciel appelé génériquement Système à Base de Connaissances (SBC).

2.2.4. Modélisation cognitive

Du point de vue de l'intelligence artificielle, la nécessité de structurer la connaissance se fait de plus en plus ressentir. Ainsi sont apparues plusieurs méthodes originales dites d'ingénierie de la connaissance ou Génie Cognitif. Ces méthodes, bien que nouvelles, ont déjà fait preuve d'efficacité dans la conception opérationnelle des SBC. Les plus connues d'entre elles, telles que KADS [Hic 89] ou KOD [Vog 89] sont actuellement de plus en plus utilisées en Europe.

Pour la réalisation de ce projet, nous avons utilisé la méthode MOISE (Méthode Organisée pour l'Ingénierie des Systèmes Experts) mise au point par J.-L. Ermine [Erm 93]. Il s'agit d'un atelier logiciel supportant une méthodologie d'ingénierie de connaissances basée sur les mêmes concepts que KADS ou KOD. Dans MOISE, la spécification formelle de la connaissance est réalisée grâce à un langage approprié. La partie de ce langage utilisée dans ce projet se divise en deux composants :
La **connaissance statique :**

Il s'agit de la modélisation du domaine de la connaissance indépendamment de l'utilisation qui en est faite. Elle spécifie la connaissance en termes d'objets et les relie entre eux par des liens fortement sémantisés grâce à l'utilisation des réseaux sémantiques. Les principaux liens utilisés sont, les liens de définition (le lien ATO, pour "ATtribute Of" ou Attribut de) et les liens classificatoires (le lien AKO, pour "A Kind Of" ou Sorte de).

La **connaissance dynamique :**

C'est une représentation de la stratégie de l'expert quand il résout la tâche qui lui incombe. Le langage formel utilisé pour la modélisation de cette connaissance se base sur une description ergonomique des tâches cognitives [Sca 89].

La spécification formelle de la connaissance peut se faire soit par un langage mathématique formel soit par un langage graphique (cf. par exemple [Alk 94]).

2.3. La base de connaissance

2.3.1. Introduction

Le travail de modélisation cognitive (ou construction de la base de connaissances) a été mené en collaboration avec Mr C. Belair qui est à la fois, secrétaire général de JD, expert juridique et l'un des concepteurs de la base documentaire JD. La première étape de cette tâche de Génie Cognitif était un travail d'extraction de la connaissance. Il s'agit de dégager la connaissance contenue dans les abstracts et impliquée dans leur création. Cette connaissance peut ne pas avoir de caractère juridique.

2.3.2. Les types de connaissance dans JD

Comparée à plusieurs autres bases documentaires juridiques (Cour de Cassation, CELEX, etc.), JD est considérée comme l'une des bases les plus structurées. En effet, un document JD (ou abstract) est une succession de champs bien définis, où chacun possède sa propre signification et son propre apport informationnel. La connaissance exprimée par ces différents champs peut être vue sous trois angles :

- Les connaissances juridiques : Regroupant des documents de jurisprudence, de la doctrine et des réponses ministérielles couvrant l'ensemble des branches du droit, la première vocation de JD est de rendre compte des différents problèmes juridiques rencontrés. Selon les concepteurs de JD, une structuration des documents en un nombre raisonnable de champs génériques est la solution la mieux adaptée pour énoncer des connaissances juridiques. Parmi ces champs on peut citer: 1) le champ indiquant la juridiction qui a rendu la décision, son siège et sa formation (intitulé et numéro de la chambre) ; 2) le champ qui résume le contenu juridique du document primaire, c'est le champ le plus informationnel de l'abstract; structuré en paragraphes, ce champ traduit l'analyse juridique faite à partir des SA ; 3) le champ qui contient un petit résumé en langage naturel décrivant brièvement le contenu factuel du document primaire correspondant.

- Les connaissances documentaires : En ce qui concerne ce genre de connaissances, on peut dire que JD ne dépasse en rien les autres bases documentaires. On retrouve des informations typiques contenues dans des champs particuliers telles que: la date de création de l'abstract, la référence du document primaire, etc.

- Les connaissances heuristiques : On entend par connaissance heuristique tout genre de connaissances exprimant un savoir faire ou une expertise propre aux concepteurs de JD. Quand on s'intéresse à la manière dont les SA on été classifiées, on s'aperçoit que cette classification n'est pas purement juridique (par purement juridique on entend la classification du législateur), elle se base en partie sur de la connaissance heuristique. En effet, le découpage de la science juridique adopté par JD est en général arbitraire sauf sur quelques matières connues et non ambiguës. Cette classification, exprimée par l'ensemble des 30 matières juridiques est construite selon plusieurs critères :

- S'inspirer de la classification légale donnée par le législateur dans un code (code civil, code de commerce, code pénal, etc.)
- Tenir compte de la doctrine qui traite des sujets n'ayant pas d'existence légale (droit des affaires, droit de l'environnement, etc.)
- Puisqu'il s'agit d'une classification des décisions judiciaires, une matière juridique n'est jamais définie a priori mais plutôt sur des critères quantitatifs et qualitatifs des documents (quantité et nature des textes présents sur la base)
- Pour des critères d'efficacité, il est inutile de donner des matières trop générales (génératrices de bruit). Une matière doit être assez spécifique afin de mieux analyser les problèmes juridiques qui lui sont affectés.

C'est l'ensemble de ces 30 matières juridiques qui constitue le premier niveau de la classification. Ce niveau a subi un peu plus d'affinement. En effet, un découpage à l'intérieur des matières juridiques a été entrepris. Chaque matière est raffinée par un ensemble de SA, d'autant plus nombreuses que le thème abordé est complexe. Ces

SA, présentant un chaînage hiérarchisé de concepts, permettent de fixer le paysage juridique dans ses moindres détails.

2. 3.3. Spécification de la connaissance statique

Une modélisation efficace et cohérente de la connaissance contenue dans les documents JD se doit de respecter les principes suivants:

- Identifier des objets qui représentent des concepts présents d'une manière explicite ou implicite dans les documents.
- Dégager les concepts pertinents pour la recherche documentaire: d'une part des concepts permettant de classifier les documents de plusieurs manières et selon plusieurs points de vue, et d'autre part des concepts juridiques extraits à partir des SA.
- Dégager des concepts factuels pertinents et essayer de les rattacher à des concepts juridiques.

Compte tenu des principes énoncés précédemment, et conformément à la méthode MOISE, un modèle cohérent de représentation des connaissances est indispensable. Pour la spécification de cette connaissance statique, on a opté pour les réseaux sémantiques (cf. § 2.2.4).

Un modèle représentatif et complet de la connaissance statique se construit progressivement. On commence par construire pour chaque concept identifié le modèle correspondant. La structuration dans un réseau sémantique du concept "Juridiction" est donnée en exemple par la figure 2.2.

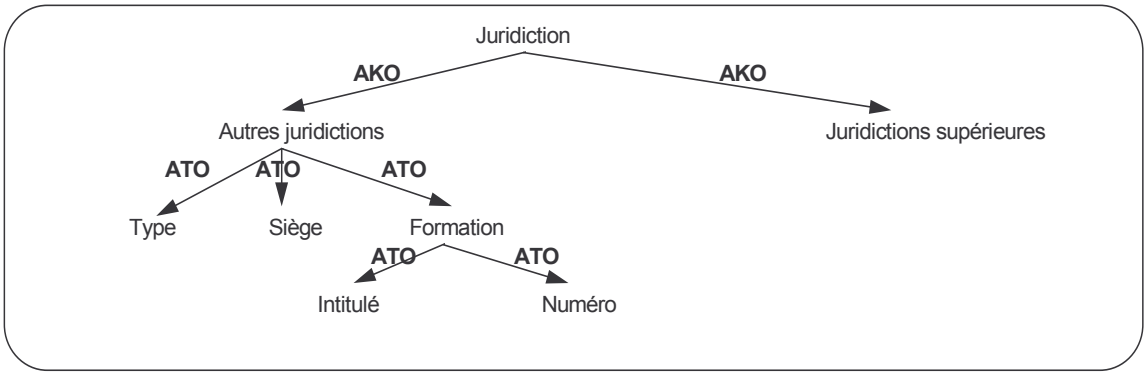

Figure 2.2 : Un exemple de réseau sémantique de la base de connaissance statique

Le réseau sémantique de la figure 2.2 peut être lu de la manière suivante : On distingue deux types de juridictions : les juridictions supérieures (Cour de Cassation, Conseil d'État) et les autres. Ces dernières sont caractérisées par leur type (Tribunal de grande instance, tribunal d'instance, etc.), leur siège (Paris, Bordeaux, etc.) et leur formation. Une formation se caractérise à son tour par l'intitulé de la chambre ayant rendu la décision (chambre civile, chambre de commerce, etc.) et son numéro.

La classification des SA a été elle aussi structurée dans un réseau sémantique. Faute de place, ce réseau ne sera pas exposé dans cet article. Celui-ci constitue sans doute l'ossature de la base de connaissance statique.

En procédant de la sorte pour chaque concept identifié, on finit par construire tout un réseau de connaissances. C'est ainsi qu'on parvient à bâtir la base de connaissance statique.

2.3.4. Spécification de la connaissance dynamique

Dans MOISE, spécifier la connaissance dynamique consiste à modéliser, moyennant un langage formel, la tâche d'un utilisateur mis au travail. Pour JD, contrairement à l'approche de l'ergonomie cognitive [Sca89], le modèle de la connaissance dynamique n'a pas été construit en observant le comportement d'un simple utilisateur, mais plutôt celui d'un expert (C. Belair). Ce choix de décrire la stratégie de l'expert, a été entrepris dans le but

d'optimiser l'utilisation de la connaissance identifiée dans la base de connaissance statique et de fournir aux utilisateurs une véritable interface "intelligente" d'aide à la recherche d'informations.

La spécification de la connaissance dynamique s'est faite grâce à un langage formel dit "langage de tâches" et suivant une analyse de type descendante. Le point de départ est la tâche complète à accomplir par le système. Cette tâche est ensuite décomposée en sous-tâche, elles mêmes décomposées en d'autres sous-tâche et ainsi de suite jusqu'à aboutir à un arbre de tâches où les feuilles sont des tâches terminales vouées à assurer des interactions spécifiques avec l'environnement (poser une question, retrouver un ensemble de documents, etc.). Chaque noeud de l'arbre est une tâche de planification de la stratégie de l'expert pour résoudre le problème au quel il est confronté (dans notre cas c'est la recherche documentaire). On distingue quatre types de tâches: La tâche d'ordonnancement, décomposée en plusieurs sous-tâche qui doivent être exécutées successivement dans l'ordre de déclaration; La tâche conditionnelle, décomposée en plusieurs sous-tâche aux quelles sont rattachées des conditions d'activation; La tâche répétitive, décrite par une tâche générique et une liste ordonnée d'objets constituant les instanciations successives du paramètre en entrée de la tâche générique; La tâche Tant que, similaire à la boucle "while...loop" dans les langages de programmation classique.

Un exemple pratique de la spécification de la connaissance dynamique est donné par la figure 2.3 (Une spécification complète de la base de connaissance dynamique nécessite une dizaine de pages de représentation graphique).

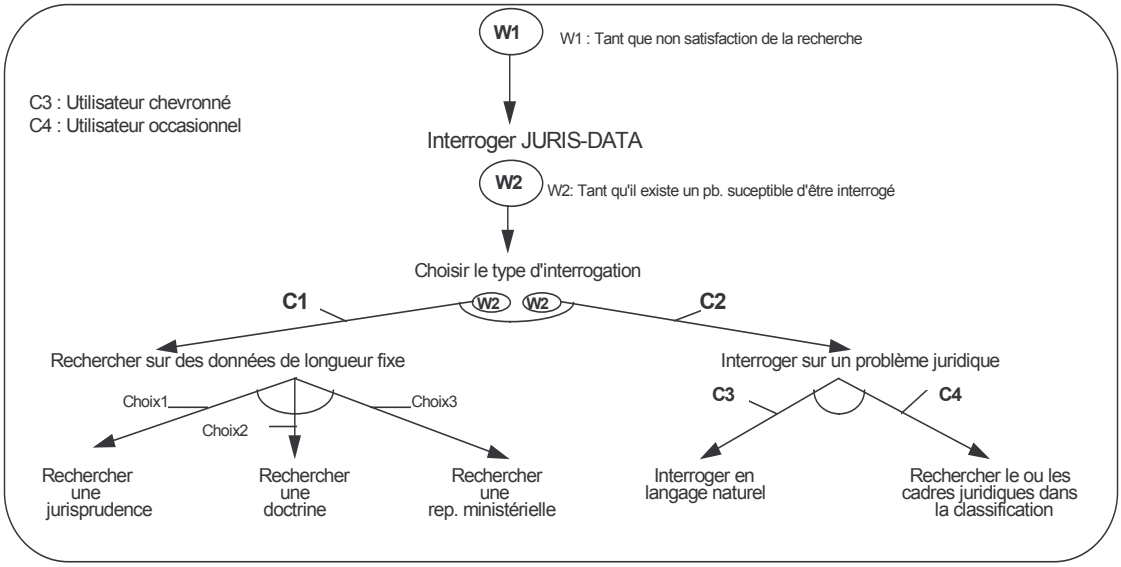

Figure 2.3 : Arbre de la tâche "Interroger JURIS-DATA"

Une brève description de cet arbre de tâche est la suivante :

Pour interroger JD, l'interface conçue propose deux options: soit une interrogation portant sur les données de longueur fixe (telles que la date de la décision, les noms des parties impliquées, le type de la juridiction ayant rendu la décision, etc.), soit une recherche d'un ou plusieurs problèmes juridiques particuliers.

En ce qui concerne la première option, l'utilisateur, selon qu'il souhaite rechercher de la jurisprudence, de la doctrine ou des réponses ministérielles, dispose pour la formulation de sa requête, de trois grilles de saisie différentes. Chaque grille lui permet de saisir des informations relatives au type de documents désirés. Ces grilles de saisie, outre le fait qu'elles constituent un mode interrogation assistée, jouent en plus le rôle d'un contrôleur à la fois syntaxique et sémantique quant à la justesse des requêtes documentaires. En effet, puisque en choisissant cette option la tâche de l'utilisateur est limitée à la simple saisie d'informations, et puisque c'est à la charge du système de structurer les informations saisies en requêtes, on est pratiquement sûr que celles ci sont syntaxiquement correctes. D'autres part, les champs proposés par une grille donnée étant propres à un type de documents précis, il n'est plus possible par exemple, contrairement à une interrogation en langage naturel, de

préciser dans une requête, des noms de parties alors que l'on souhaite retrouver de la doctrine; d'où le rôle de contrôleur sémantique.

Quant à la seconde option (Interroger sur des problèmes juridiques), l'utilisateur, suivant qu'il est chevronné ou occasionnel, dispose à ce niveau également de deux types d'interrogations:

- Une interrogation libre en langage naturel qui suppose que, ce dernier étant chevronné, peut sans aucune aide de la part du système parvenir à formuler correctement une requête documentaire.

- Une interrogation assistée, destinée généralement aux utilisateurs novices, qui consiste à les aider à retrouver, en se basant sur la classification des SA, le ou les cadres juridiques qui décrivent le mieux, voire exactement, leurs problèmes. L'idée de base qui est à l'origine de la mise en place de ce mode assisté d'interrogation est la suivante: Puisque tout abstract sur JD est généré à partir d'une SA bien précise, et puisque l'on dispose déjà d'une classification des SA, on peut, en rattachant les abstracts de JD aux feuilles de la classification des SA, construire une véritable taxinomie des documents de la base; d'ailleurs c'est ce qui a été fait sans difficulté particulière. La spécification d'une telle interface est donnée par la figure 2.4.

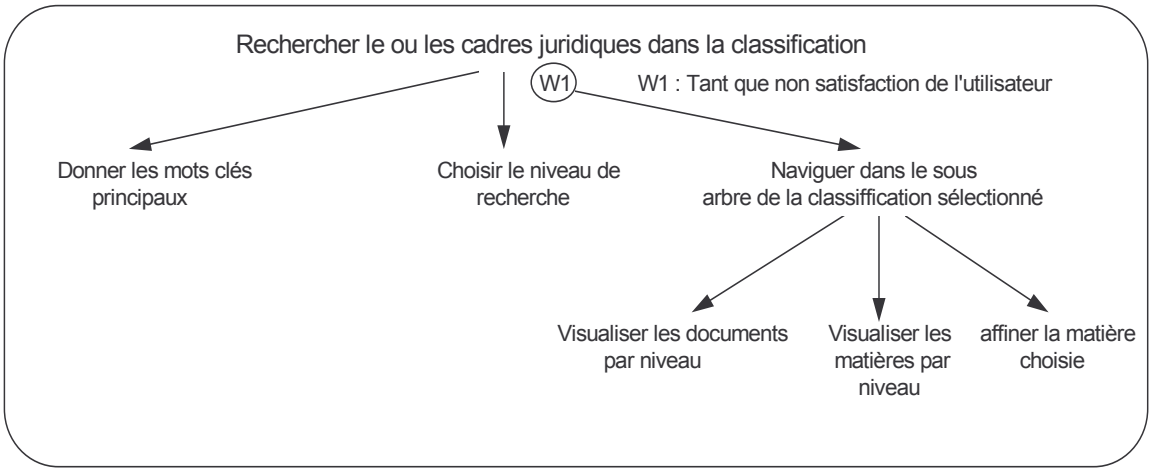

Figure 2.4 : Arbre de la sous-tâche "Rechercher le ou les cadres juridiques dans la classification"

Ayant une petite idée sur le type des problèmes juridiques recherchés, l'utilisateur commence par donner jusqu'à trois mots clés jugés significatifs. Le système génère alors toutes les requêtes possibles par combinaison de ces mots. La recherche est alors lancée et le résultat s'affiche sous la forme d'une grille tridimensionnelle où chaque ligne contient, la combinaison des mots clés ayant servi à la recherche, le nombre de documents trouvés ainsi que les matières juridiques correspondantes. Celles-ci ne sont autres que les noeuds du premier niveau de la taxinomie des documents de la base. L'utilisateur peut ainsi, en fonction du nombre des documents trouvés et du type des matières juridiques proposées, orienter sa recherche en choisissant le niveau de recherche qui lui parait le plus pertinent. Un niveau de recherche correspond à l'ensemble {combinaison de mots clés, nombre de documents correspondants, matières juridiques correspondantes}.

Une fois le niveau de recherche choisi, l'utilisateur peut affiner sa recherche en explorant la branche de la classification couvrant la matière juridique sélectionnée. A chaque sous niveau de la classification, une nouvelle requête plus précise est générée et la recherche et de nouveau relancée. L'utilisateur peut arrêter son exploration de la taxinomie quand le nombre des documents trouvés devient assez raisonnable pour permettre leur consultation et par conséquent le choix des plus pertinents d'entre eux. Cette navigation dans la classification traduit en fait le cheminement de l'esprit d'un expert lors de la formulation d'une requête documentaire.

2.3.5. Conception, codage et validation

La première interface conçue à partir de cette spécification était un programme générique d'instanciation de réseaux sémantiques. Un module reçoit en entrée la spécification formelle d'un concept juridique donné pour générer en sortie le réseau sémantique correspondant. Formuler une requête revient en gros à instancier un ou

plusieurs réseaux sémantiques grâce à un module récursif de parcourt d'arbre. Cette interface était estimée trop sophistiquée et incompréhensible par les utilisateurs de JD.

Une seconde interface plus souple et plus abordable a été ensuite réalisée. L'arbre de tâche de la connaissance dynamique a été fidèlement codé en algorithmes. La souplesse d'utilisation de l'interface repose précisément sur l'utilisation de procédés simples: menus à choix numériques, grilles de saisie, quelques touches de fonctions, etc. Outre l'efficacité, l'objectif d'une telle interface est d'être le plus simple possible et ce pour deux raisons: d'une part, les juristes ne sont pas forcément habitués aux interfaces graphiques sophistiquées (telles que les interfaces Windows) ; d'autre part, un tel logiciel est destiné à fonctionner dans un environnement restreint et non graphique : le Minitel Français.

Cette interface a été développée en langage C sous UNIX. Elle a été testée sur une partie significative de JD (20% environ).

Pour la validation de notre interface, nous avons entrepris une démarche en deux phases: une validation de la complétude et une validation de la justesse.

Complétude de l'interface : Le but de cette validation est de montrer que toutes les tâches et sous-tâches de la connaissance dynamique sont effectivement réalisées par l'interface. Elle consiste à établir une correspondance bijective entre les différentes sous-tâches constituant l'arbre général et les différents écrans et fonctions proposés par l'interface.

Justesse de l'interface : Il s'agit de s'assurer de l'adéquation des requêtes produites par le système par rapport aux problèmes posés. Pour manque de méthodes formelles connues de vérification de justesse, nous avons conçu notre propre méthode. Celle-ci consiste à évaluer, pour chaque type d'utilisateurs, la justesse des requêtes obtenues grâce à l'interface. Cette évaluation repose sur la confrontation des requêtes obtenues aux requêtes de référence déterminées par l'expert. Les résultats des tests effectués sont résumés dans des fiches de validation dont le modèle est donné par la figure 2.5.

Acteur Action	Expert	Initié	Néophyte
Problème juridique			
Requête libre			
Requête assistée			

Figure 2.5 : Fiche de validation

Les colonnes représentent les trois types d'utilisateurs : Expert dans JD, utilisateur initié et utilisateur néophyte. La première ligne traduit en langage naturel le problème juridique à résoudre, la seconde exprime la requête formulée par l'utilisateur en utilisant l'option "interrogation libre" de l'interface et la troisième présente la requête obtenue suite à l'utilisation du mode assisté. Il importe de noter que chaque fiche de validation comporte, outre les requêtes obtenues par l'utilisateur en guise de réponse au problème juridique posé, une requête de référence donnée par l'expert (Mr C. Belair).

Un nombre raisonnable de fiches de validation dûment remplies a globalement prouvé la justesse de l'interface.

3. Généralisation de l'interface d'interrogation pour d'autres bases documentaires juridiques non structurées

On a vu que, l'interface de recherche d'informations décrite précédemment s'appuie sur une base de connaissances construite essentiellement à partir d'une classification des documents des juristes. Il est donc indispensable, afin de pouvoir la transposer sur d'autres bases documentaires juridiques, de disposer des taxinomies correspondantes. Malheureusement, certaines bases documentaires juridiques, notamment celles basées sur le texte intégral, ne possèdent aucune structure classificatoire. Par conséquent, il importe, si on veut généraliser l'utilisation de l'interface pour ce genre de bases documentaires, de déterminer une méthodologie organisée de construction de structures classificatoires à partir de documents juridiques. Assurément, Il existe des outils dédiés à accomplir ce genre de tâches, comme le système CABARET [Ska 89], mais, notre désir de formaliser le savoir faire de l'expert et de maîtriser l'outil nous a conduit à développer notre propre méthode.

3.1. Méthodologie de construction de structures classificatoires

3.1.1. Introduction

La finalité de cette partie du travail est de définir une méthodologie de construction de classifications juridiques en s'appuyant sur le savoir faire d'un expert. En effet, construire avec l'expert une taxinomie du domaine de référence constitue un moyen sûr et commode d'approcher la structure qui sera donnée à la base de connaissances [Vog 88]. De plus, l'explicitation et la formalisation de la stratégie adoptée par l'expert lors de la résolution de cette tâche classificatoire constitue une véritable méthodologie organisée de construction de structures classificatoires.

Dans la suite de cet article, nous nous contenterons d'exposer les différentes étapes de la méthode. Il est tout de même intéressant de noter que celle-ci n'est autre que la spécification formelle de la démarche entreprise par un expert juridique (C. Belair) pour construire une classification. L'exemple exposé ici correspond à la classification rattachée au titre VI (du divorce), du livre premier (des personnes) du code civil français (Art 229 à 310).

Avant d'aller plus loin, il convient de donner quelques définitions :

- *Lexème primaire*: terme dans le sens ne peut être déduit de celui de ces composants.
- *Lexème secondaire*: composé de plusieurs lexèmes primaires. C'est, par exemple, une association entre un nom et un ou plusieurs adjectifs.
- *KWIC*: **K**ey **W**ord **I**n **C**ontext : c'est un ou plusieurs termes pris dans le texte. Un KWIC peut être un simple mot, un lexème primaire, un lexème secondaire, une phrase complète ou un bout de phrase.
- *KWOC*: **K**ey **W**ord **O**ut of **C**ontext : c'est un synonyme ou analogue d'un KWIC du texte.
- *KWAC*: **K**ey **W**ord **A**t **C**ontext : c'est l'explicitation d'un concept implicite.

3.1.2. Les différentes étapes de la méthode

Afin d'illustrer les différentes étapes de la méthode par un exemple complet, nous allons considérer certains articles du code civil français à savoir: Art1382, 1383, 1384, 1385 et 1386. Ces articles correspondent au chapitre II : *Délits et quasi-délits*, du titre IV : *Des engagements qui se forment sans convention*, du livre III : *Des différentes manières dont on acquiert la propriété*.

⇒ Etape 1: Extraction des KWIC

→ **Entrée :** Unités documentaires juridiques à classifier
← **Sortie :** Unités documentaires juridiques réduites aux KWIC extraits

L'expert commence par scruter les Unités documentaires[*] afin d'identifier les KWIC jugés pertinents pour la construction de la classification. Cette première étape repose essentiellement sur la faculté de l'expert de reconnaître les KWIC résumant en quelque sorte l'intégralité de l'unité documentaire. Le résultat de cette première étape appliquée sur les articles choisis du code civil est donné par la figure 3.1, les KWIC identifiés par l'expert sont soulignés.

⇒ Etape 2: Détermination des KWOC

→ **Entrée :** Unités documentaires juridiques réduites aux KWIC extraits
← **Sortie :** Unités documentaires juridiques réduites aux KWIC extraits et KWOC correspondants

Pour chaque KWIC identifié lors de l'étape 1, l'expert détermine le ou les KWOC correspondants. Pour ce faire, il peut, soit s'appuyer sur ses propres connaissances du vocabulaire juridique, soit se référer à un thesaurus. L'expérience prouve que plusieurs expressions clés peuvent avoir les mêmes KWOC.

[*] Pour cet exemple, l'unité documentaire est un article du code civil Français

Art 1382: Tout fait quelconque de l'homme, qui cause à autrui un dommage, oblige celui par la faute duquel il est arrivé à le réparer.

Art 1383: Chacun est responsable du dommage qu'il a causé non seulement par son fait, mais encore par sa négligence ou par son imprudence.

Art 1384: **§1** On est responsable non seulement du dommage que l'on cause par son propre fait, mais encore de celui qui est causé par le fait des personnes dont on doit répondre, ou des choses que l'on a sous sa garde.
§2 toutefois, celui qui détient, à un titre quelconque, tout ou partie de l'immeuble ou des biens mobiliers dans lesquels un incendie a pris naissance ne sera responsable, vis-à-vis des tiers, des dommages causés par cet incendie que s'il est prouvé qu'il doit être attribué à sa faute ou à la faute des personnes dont il est responsable.
Cette disposition ne s'applique pas aux rapports entre propriétaires et locataires, qui demeurent régis par les articles 1733 et 1734 du code civil.
§3 Le père et la mère, en tant qu'ils exercent le droit de garde, sont solidairement responsable du dommage causé par leurs enfants mineurs habitant avec eux.
§4 Le maître et les commettants, du dommage causé par leurs domestiques et préposés dans les fonctions auxquelles ils les ont employés.
§5 Les instituteurs et les artisans, du dommage causé par leurs élèves et apprentis pendant le temps qu'ils sont sous leur surveillance.
§6 La responsabilité ci-dessus a lieu, à moins que les père et mère et les artisans ne prouvent qu'ils n'ont pu empêcher le fait qui donne lieu à cette responsabilité.
§7 En ce qui concerne les instituteurs, les fautes, imprudences ou négligences invoquées contre eux comme ayant causé le fait dommageable, devront être prouvées, conformément au droit commun, par le demandeur, à l'instance.

Art 1385: Le propriétaire d'un animal, ou celui qui s'en sert, pendant qu'il est à son usage, est responsable du dommage que l'animal a causé, soit que l'animal fût sous sa garde, soit qu'il fût égaré ou échappé.

Art 1386: Le propriétaire d'un bâtiment est responsable du dommage causé par sa ruine, lorsqu'elle est arrivée par suite du défaut d'entretien ou par le vice de sa construction.

Figure 3.1 : La sélection des KWIC faite par l'expert pour les articles 1382 à 1386 du code civil français

⇒ Etape 3: Sélection des expressions clés

→ **Entrée :** Unités documentaires juridiques réduites aux KWIC extraits et KWOC correspondants
← **Sortie :** Unités documentaires juridiques réduites aux expressions clés sélectionnées
Parmi l'ensemble des KWIC et des KWOC déjà recensés, l'expert procède à la sélection des expressions clés qui seront considérées dans les étapes suivantes. Ces expressions contribuent à la détermination du vocabulaire de la classification. Quant aux expressions non sélectionnées, elles peuvent servir à la construction d'un index alphabétique. Les résultats des étapes 2 et 3, appliquées sur l'exemple précédent, sont donnés par la figure 3.2 (les KWOC déterminés par l'expert sont marqués par **(=** ....**)**, les expressions clés sélectionnées sont soulignées).

Art 1382: fait quelconque de l'homme (= fait propre); cause; à autrui un dommage (= préjudice); faute; réparer (= réparation)

Art 1383: responsable (= responsabilité); dommage (= préjudice); causé (= cause); son fait (= fait propre); négligence; imprudence

Art 1384:§1: responsable (= responsabilité); dommage (= préjudice); cause; son propre fait (= fait propre); fait des personnes dont on doit répondre (= personnes dont on doit répondre) ; choses que l'on a sous sa garde. (= chose dont on la garde)

§2: celui qui détient (= propriétaire ou locataire); immeuble; biens mobiliers (= meubles)
incendie a pris naissance (= naissance d'incendie); tiers; prouvé (= preuve); sa faute (= faute propre); faute des personnes dont il est responsable (= faute des personnes dont on est responsable); ne s'applique pas aux rapports entre propriétaires et locataires (= exclusion des rapports entre propriétaire et locataire)

§3: père (= parent); mère (= parent); droit de garde (= exercice du droit de garde); solidairement responsable (= responsabilité in solidum des parents); dommage causé (= préjudice); par leurs enfants mineurs habitant avec eux. (= cohabitation parents-enfants mineurs) (= enfants mineurs)

§4: maître (= employeur); commettants (= employeurs); dommage causé par leurs domestiques et préposés (= dommage causé par les employés) (= employés)

§5: instituteurs (= professeurs); artisans; dommage causé par leurs élèves et apprentis (= dommage causé par les élèves ou apprentis) (= élèves\apprentis); surveillance

§6: père et mère et les artisans (= parents, = patrons); ne prouvent qu'ils n'ont pu empêcher le fait (= preuve négative du fait)

§7: instituteurs (= professeurs);
fautes, imprudences ou négligences
devront être prouvées
} (=preuve positive des fautes, imprudences, négligences)

Art 1385: propriétaire d'un animal, ou celui qui s'en sert; usage (= utilisation d'un animal); responsable du dommage que l'animal a causé (= responsable du dommage causé par un animal); animal fût sous sa garde (= garde de l'animal); égaré ou échappé. (= animal échappé ou égaré)

Art 1386: propriétaire d'un bâtiment (= propriétaire d'un immeuble); responsable du dommage causé (= préjudice causé); ruine (= ruine d'un immeuble); défaut d'entretien; vice de sa construction. (= vice de construction)

Figure 3.2 : Détermination des KWOC et sélection des expressions clés

⇒ *Etape 4: Détermination des KWAC*

→ **Entrée :** Unités documentaires juridiques réduites aux expressions clés sélectionnées
← **Sortie :** Unités documentaires juridiques réduites aux expressions clés sélectionnées et aux KWAC déterminés.

Pour chaque unité documentaire juridique, l'expert détermine les KWAC qu'elle évoque. Identifier un KWAC revient, soit à expliciter un concept implicite, soit à agréger certaines expressions clés sur la base des relations qu'elles entretiennent entre elles (lien de causalité, lien de subordination, etc.).

L'agrégation des expressions clés en KWAC, consiste à reconnaître des discontinuités et/ou des similarités sémantiques. L'expert parcourt la liste des expressions et repère une variation de sens à partir d'un certain élément. Cette variation de sens correspond à une variation de trait sémantique (ou *sème*) qui lui permet d'élaborer un KWAC. La détermination d'un nouveau KWAC par l'expert, traduit une reconnaissance de critères discriminants entre les différentes expressions clés. Ce phénomène de discontinuité sémantique, identifiable par un sème particulier, est un des fondements d'une théorie du sens, appelée sémiotique (théorie du signe), c'est ce qu'on appelle une ***rupture d'isotopie***. (cf. [Gre66] ou [Erm89]). L'isotopie est la redondance d'un trait sémantique, les "effets de sens" naissent des ruptures d'isotopie. Pour donner un exemple simple, la phrase "il y a de l'orage dans l'air" possède deux sens suivant qu'on lui met par exemple le trait sémantique /humain/ ou /non-humain/. Si on lui rajoute la phrase "ils se regardent méchamment" (qui possède le trait /humain/), ou la phrase "il y a de gros nuages dans le ciel" (qui possède le trait /non-humain/), l'isotopie sémantique du trait lève l'ambiguïté du sens.

Pour chaque nouveau KWAC déterminé, l'expert revient sur les unités documentaires précédemment traitées afin de s'assurer de la complétude de son analyse. En effet, certains KWAC paraissant pertinents dans certaines unités documentaires, peuvent échapper à l'expert lors de son analyse à d'autres qui, pourtant les évoquent. Ce retour en arrière permet d'éviter ce genre d'oubli. Les résultats de cette étape sont les suivants :

Art 1382:	Responsabilité civile Lien de causalité entre son propre fait et le dommage
Art 1383:	Responsabilité civile Lien de causalité entre son propre fait et le dommage
Art 1384:	Responsabilité civile
§1:	Lien de causalité entre le dommage et autre chose Présomption de responsabilité
§2:	Chose dont on a la garde Communication d'incendie Exclusion
§3:	Personne dont on doit répondre
§4:	Personne dont on doit répondre Lien de subordination
§5:	Personne dont on doit répondre
§6:	Exonération de responsabilité
§7:	Exonération de responsabilité
Art 1385:	Responsabilité civile Chose dont on a la garde Lien de causalité entre le dommage et autre chose Présomption de responsabilité
Art 1386	Responsabilité civile Chose dont on a la garde Lien de causalité entre le dommage et autre chose Présomption de responsabilité

Figure 3.3 : Liste des KWAC déterminés par l'expert

⇒ *Etape 5: Construction des structures classificatoires*

→ **Entrée :** Unités documentaires juridiques réduites aux expressions clés sélectionnées et aux KWAC déterminés
← **Sortie :** Structures classificatoires
Pour la construction de la classification, l'expert procède par agrégation. C'est une approche ascendante, agrégeant des sous-classes en classe. Elle consiste à regrouper les entités à classifier en petits lots, ces petits lots son à nouveau regroupés en lots plus importants, et ainsi de suite jusqu'à n'obtenir qu'une seule classe générale. Ce type d'approche suppose l'utilisation d'un des principes de base de la généralisation inductive, la "règle de généralisation ascendante" énoncée par [Mic83]. Généralement, le regroupement de sous classes en classe s'effectue sur la base des relations sémantiques qu'elles entretiennent entre elles. On peut citer à titre d'exemple les deux types de relations sémantiques les plus fréquentes :

- Les relations hiérarchiques (ou **hiérarchie**) qui se scindent en deux catégories : la *généricité/spécificité* (ex. entre Divorce pour faute et Divorce), et la *partitivité* (ex. entre Nez et Tête).

- Les relations associatives(ou **association**) exprimant une relation symétrique entre deux concepts qui, bien que non liés entre eux par une hiérarchie, sont susceptibles de s'évoquer mutuellement, par associations d'idées. Les types d'associations sont assez diversifiés, signalons en exemple les cas suivants : *Causalité :* entre Maladie et Infection, *Instrumentation* : entre Auscultation et stéthoscope, *Localisation:* entre opération chirurgicale et Hôpital, etc. Pour notre exemple, la classification se construit progressivement. Tout d'abord, l'expert commence par établir des liens entre les KWIC /KWOC retenus lors de l'étape 3 et les KWAC qui les agrègent (cf. figure 3.4), toujours en s'appuyant sur le principe de rupture d'isotopie.

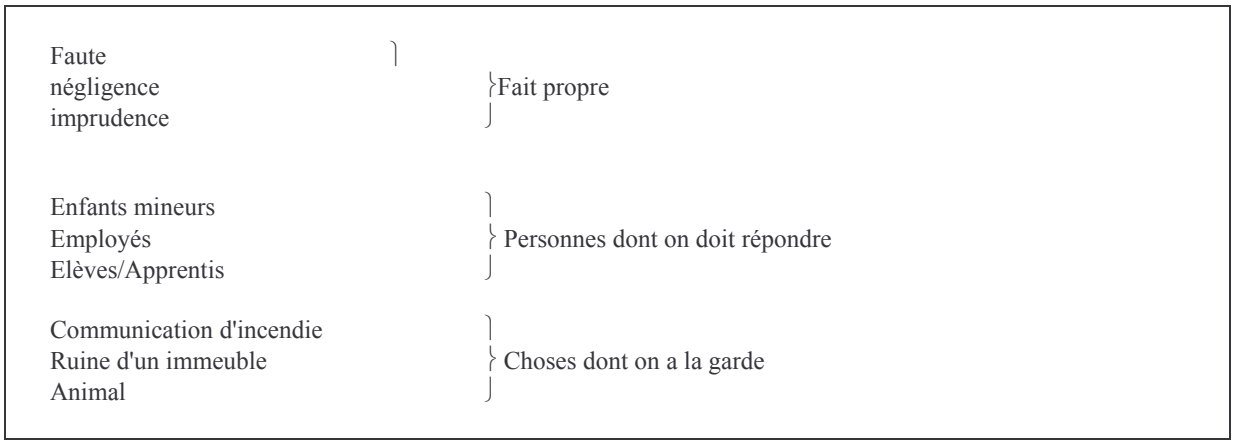

Figure 3.4 : Établissement des liens entre les KWIC/KWOC et les KWAC qui les agrègent

Ensuite, on établit les liens entre les KWAC d'un même niveau et ceux des niveaux hiérarchiquement supérieurs. Le regroupement de la figure 3.5 exprime la rupture d'isotopie entre personne et chose.

Figure 3.5 : Rupture d'isotopie entre personne et chose

Figure 3.6 : Rupture d'isotopie entre fait propre et fait d'autrui ou de la chose

Quant au regroupement de la figure 3.6, il traduit d'une part, la rupture d'isotopie fait propre et fait d'autrui ou de la chose, et d'autre part la redondance successive (mais informationnelle) du KWAC "Responsabilité civile" dans les différents articles analysés, c'est ce qui fait que ce KWAC vient coiffer tous les concepts de la classification.

Il importe de noter que lors de cette étape l'expert peut ne pas considérer dans la classification finale certaines entités inclassables ou ayant un faible apport informationnel. De telles entités contribueront à la construction de l'index. Il est aussi tolérable d'affiner la classification par la détermination de nouveaux KWAC afin d'exprimer une rupture d'isotopie omise lors des phases précédentes.

La classification finale de ces articles est donnée par la figure 3.7. C'est à partir de cette classification que sera construite la base de connaissances sur laquelle s'appuie l'interface de recherche d'information décrite précédemment.

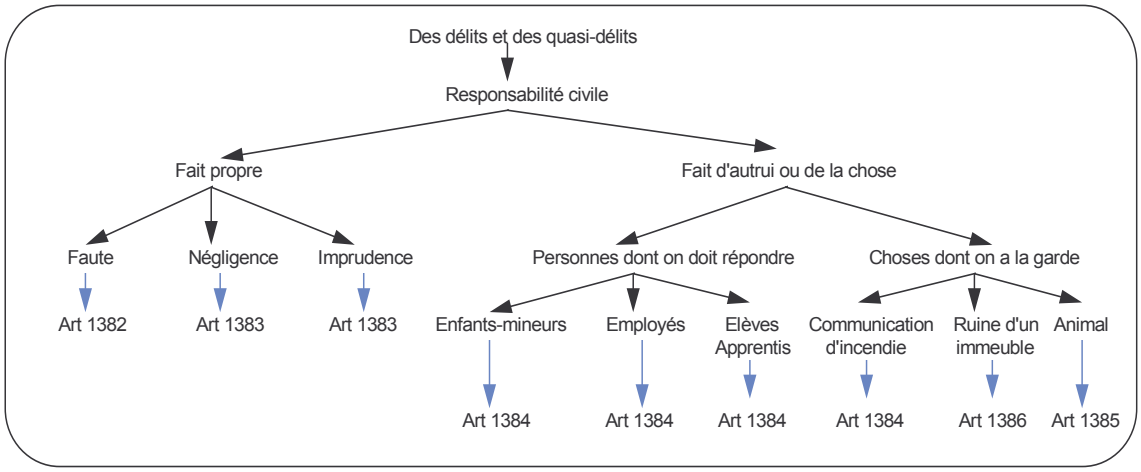

Figure 3.7 : Classification du Chapitre II du Titre IV du Livre III du code civil français

3.2. Méthodologie de construction d'index

3.2.1. Utilité des index

Comme la méthode de construction de structures classificatoires parvient à fournir un ensemble complet des concepts juridiques pertinents et présents (d'une manière implicite ou explicite) dans les documents constituant la base, ainsi que leurs synonymes, analogues, etc., il est possible d'exploiter cette abondance de vocabulaire afin de construire des index alphabétiques. Une autre méthodologie organisée de construction d'index a donc été développée, elle sera brièvement exposée dans la suite de cet article.

De tels index alphabétiques peuvent avoir une double finalité :

* Servir de thesaurus pour l'enrichissement et/ou l'indexation des documents. En effet, l'établissement de liens entre les concepts qu'évoque un document et ceux de l'index permet l'accès à l'information par différents termes possibles.

* Ces index, se présentant comme les index "papier" qu'on retrouve à la fin de chaque livre, offrent à l'utilisateur une autre option de recherche documentaire, qui pour certaines bases juridiques (telles que le code civil ou le code du travail), est très intéressante. La mise en ligne des index, permet à l'usager, en tapant le concept recherché (ou l'un de ses synonymes, analogues, etc.) de retrouver directement les références des articles qui l'évoquent. Il s'agit du même principe que la recherche dans un index "papier".

3.2.2. Choix de la structure des index

Généralement, un index alphabétique structuré comporte des unités lexicales et des relations sémantiques entre ces unités. On dénombre quatre unités lexicales :

- Les **descripteurs**. L'AFNOR définit un descripteur comme un "mot ou groupe de mots retenus dans un thesaurus et choisis parmi un ensemble de termes équivalents pour représenter sans ambiguïté une notion contenue dans un document ou dans une demande de recherche documentaire." Un descripteur est donc un élément représentatif d'une classe d'équivalence composée de plusieurs termes rattachés au même concept.

- Les **non descripteurs** définis selon l'AFNOR comme étant des "mots ou groupes de mots figurant dans un thesaurus avec interdiction d'emploi et renvoi à un ou plusieurs descripteur à utiliser." Ce sont des synonymes du descripteur, les autres éléments de la classe d'équivalence.

- Les **mots-outils** sont des mots presque toujours unitermes (isolés), de sens relativement peu précis et non inclus dans un thème spécifique. Ils sont listés, sans relation sémantique ni même de renvoi vers d'autres termes.

- Les **groupes de descripteurs** qui rassemblent chacun quelques dizaines de descripteurs.

Dans un index, ces unités lexicales peuvent être structurées de plusieurs manières, on distingue :

- Les index à structure "linéaire" : L'index se présente sous forme d'une liste d'unités lexicales souvent considérées comme des non descripteurs ou des mots-outils et triée par ordre alphabétique. Après chaque unités lexicales on inscrit, soit un ou plusieurs renvois aux unités documentaires qui l'évoquent, soit un renvoi au descripteur correspondant.

- Les index à structure "hiérarchique" : L'index est constitué de petits bouts de classification. Chaque entrée de l'index agrège un ensemble de concepts qui à leur tour peuvent agréger d'autres sous ensemble. Les renvois vers les unités documentaires apparaissent au niveau le plus profond de la hiérarchie.

- Les index à structure "hybride" : Ils intègrent les deux structures précédentes à savoir la structure "linéaire" et la structure "hiérarchique".

Pour la construction de nos index alphabétiques juridiques, on a opté pour la structure "hybride". Ce choix qui est loin d'être arbitraire est justifié par les raisons suivantes :

* Une structure "linéaire" ne parvient pas souvent à résoudre le problèmes des mots polysèmes qui doivent forcément apparaître dans leur contexte; d'où la nécessité d'une structure "hiérarchique".

* Dans le code civil français par exemple, on retrouve des mots autonomes et spécifiques à certains articles. Ces mots n'appartenant à aucune hiérarchie requièrent un index à structure "linéaire".

3.2.3. Les différentes étapes de la méthode

Nos deux méthodologies de construction de structures documentaires (structures classificatoires et index alphabétiques) présentent certaines étapes en commun. Il s'agit des étapes 1 à 4 (voir § 3.1.3) qui consistent à déterminer le vocabulaire exprimant l'ensemble des concepts juridiques traités. Une fois ce vocabulaire juridique réuni, la méthodologie de construction d'index alphabétiques propose d'autres étapes ayant pour finalité la structuration de ce vocabulaire en index. Seules ces dernières seront exposées en détails.

⇒ *Etape 1: Extraction des KWIC* (Idem que pour la construction des structures classificatoires)
⇒ *Etape 2: Détermination des KWOC* (Idem que pour la construction des structures classificatoires)
⇒ *Etape 3: Sélection des expressions clés*

Il s'agit de la même étape que pour la construction des structures classificatoires mais avec des critères différents pour la sélection des expressions. En effet, plus notre index est riche en vocabulaire plus il est complet et efficace. Il est donc important de tout sélectionner sauf les KWIC ayant une forme inadaptée pour figurer dans un index (formes verbales, bouts de phrases, etc.).

⇒ *Etape 4: Détermination des KWAC* (Idem que pour la construction des structures classificatoires)
⇒ *Etape 5: Tri alphabétique des KWIC, KWOC et KWAC suivis par les références des unités documentaires*
→ **Entrée :** Unités documentaires juridiques réduites aux expressions clés sélectionnées et aux KWAC déterminés
← **Sortie :** Liste alphabétique des KWIC, KWOC et KWAC suivis par les références des unités documentaires qui les évoquent

On commence par rattacher à chaque expression la référence de l'unité documentaire qui l'évoque. Ces expressions sont ensuite triées par ordre alphabétique croissant.

⇒ *Etape 6: Établissement des liens "VOIR"*

→ **Entrée :** Liste alphabétique des KWIC, KWOC et KWAC suivis par les références des unités documentaires qui les évoquent
← **Sortie :** Partie à structure "linéaire" de l'index

Il s'agit de retenir les descripteurs et les non descripteurs (synonymes, analogues proches, allotaxies et éventuellement les antonymes) et les rattacher par le lien "VOIR" (cf. figure 3.8).

Préjudice	**Voir**	Dommage
Réparer	**Voir**	réparation

Figure 3.8 : Exemple de lien "VOIR"

⇒ *Etape 7: Agrégations de concepts sous des facteurs communs*
→ **Entrée :** Unités documentaires juridiques réduites aux expressions clés sélectionnées et aux KWAC déterminés
← **Sortie :** Agrégations de concepts sous les facteurs communs

On commence par repérer les unités lexicales présentant un même facteur commun. Par facteur commun, on entend un mot ou un groupe de mots présents explicitement dans plusieurs unités lexicales, Il peut être soit un descripteur soit un mot-outil Les unités lexicales repérées précédemment sont en suite regroupés sous le facteur commun identifié (cf. figure 3.9). Lors de cette étape, on se réserve le droit de ne pas considérer certains facteurs du fait de leur maigre apport informationnel dans le domaine juridique.

Dommage:	Dommage	Art 1382
		Art 1383
		Art 1384
		Art 1385
		Art 1386
	Dommage causé par les élèves ou apprentis	Art 1384
	Dommage causé par les employés	Art 1384
	Lien de causalité entre le dommage et autre chose	Art 1384
		Art 1385
		Art 1386
	Lien de causalité entre son propre fait et le dommage	Art 1382
		Art 1383

Figure 3.9 : Exemple d'agrégation sous un facteur commun

⇒ *Etape 8: Affectation des termes résiduels*

→ **Entrée :** Unités documentaires juridiques réduites aux expressions clés sélectionnées et aux KWAC déterminés
← **Sortie :** Index non complètement structuré

Par termes résiduels on désigne les unités lexicales n'ayant pas été retenues lors des deux étapes précédentes. Ces termes sont tout d'abord listés afin qu'on puisse les caser sous les concepts qui les subordonnent. Quant aux termes restants (que l'on n'arrive pas à caser), ils figureront en tant qu'entrées autonomes dans l'index.

⇒ *Etape 9: Classification de ce qui est classifiable sous les facteurs communs*

→ **Entrée:** Agrégations de concepts sous les facteurs communs
← **Sortie :** Index non complètement structuré

Il s'agit d'affiner la hiérarchisation de l'index en déterminant d'autres facteurs communs plus complexes.

Dommage:	Dommage	Art 1382

		Art 1383
		Art 1384
		Art 1385
		Art 1386
Dommage causé par:	les élèves ou apprentis	Art 1384
	les employés	Art 1384
Lien de causalité entre:	le dommage et autre chose	Art 1384
		Art 1385
		Art 1386
	son propre fait et le dommage	Art 1382
		Art 1383

Figure 3.10 : Exemple de classification de ce qui est classifiable sous les facteurs communs

⇒ *Etape 10: Organisation de l'index et vérification de cohérence*

→ **Entrée :** Index non complètement structuré
← **Sortie :** Index alphabétique

C'est la dernière étape de la méthode. Elle consiste en un tri alphabétique des différentes entrées de l'index suivi d'une vérification de complétude et de cohérence. Après cette dernière étape, l'index est prêt à l'exploitation.

3.3. Automatisation de certaines étapes des deux méthodes

Comme on a pu le constater, la mise en pratique des deux méthodologies de construction de langages documentaires, repose essentiellement sur l'intervention d'un expert juriste. Cette approche à caractère fondamentalement "humain", constitue un moyen commode et sûr d'extraire toute la connaissance sur le domaine et de la structurer en structures documentaires. En effet, c'est grâce aux connaissances juridiques de l'expert qu'on arrive par exemple à identifier des concepts implicites ou à mesurer la pertinence d'un concept par rapport à la matière juridique traitée, ce qui n'est pas à notre avis parfaitement réalisable par des automatismes, contrairement à ce qu'explique G. Salton dans [Sal 72].

Dans le cas où aucun expert n'est disponible, diverses techniques automatiques d'extraction et de structuration de connaissances peuvent être utilisées. Ces techniques, bien qu'elles ne puissent pas remplacer l'expert sans perte de performances, arrivent souvent à fournir des résultats assez satisfaisants. Nous proposons dans la suite de cet article quelques unes de ces techniques.

3.3.1. Extraction automatique de KWIC

Pour l'automatisation de cette tâche qui constitue la première étape de nos deux méthodes, on propose d'exploiter certains outils d'acquisition d'unités terminologiques utilisés généralement par les systèmes d'indexation automatique.

Le système SPIRIT

Le système SPIRIT [SPI 92] produit automatiquement un thesaurus et une liste de mots fonctionnels, et assure la recherche de documents [And 83a], [And 83b]. L'extraction de la terminologie utilisée dans l'élaboration du thesaurus fait appel à des méthodes linguistiques :

- L'analyse morphologique recherche pour chaque mot son identité lexicale ainsi que ses valeurs grammaticales possibles.
- L'analyse grammaticale lève les ambiguïtés non résolues au niveau morphologique.
- Un filtrage morphosyntaxique extrait du corpus toutes les configurations linguistiques correspondant à une morphologie ou à une structure syntaxique donnée. Ses résultats son ensuite manuellement filtrés.

TERMINO [DAV]

Le système TERMINO n'utilise pas de connaissances lexicales (il n'y a pas de dictionnaire). Cette approche est basée sur l'analyse de la structure interne et externe des groupes de mots potentiellement intéressants.

Les unités terminologiques recherchées sont appelées des synapsies. Il s'agit d'unités polylexicales caractérisées par une structure hiérarchiques interne et qui occupent une position noyau en tant que groupe nominal. Pour les identifier, le système TERMINO procède à une analyse morphologique qui lemmatise et catégorise dynamiquement les termes. Le module syntaxique construit des arbres d'expressions parenthèsées et emboîtées. Les expressions candidates sont filtrées par application d'heuristiques pour isoler les synapsies correctes.

Cette méthode est performante et est actuellement commercialisée mais elle demande de grands volumes de connaissances explicites : grammaire et ensemble d'heuristiques.

LEXTER [Bou 92]

Le système LEXTER recherche les unités terminologiques assimilées à des groupes nominaux composées figés dont la structure grammaticale est particulière. Comme TERMINO, il s'appuie sur une analyse des structures interne et externe des groupes nominaux quoique l'analyse syntaxique y soit complète. Il se distingue par sa stratégie descendante : il s'agit de produire le plus grand nombre d'unités terminologiques potentielles puis de les filtrer.

L'analyse morphologique établit les catégories grammaticales des termes et résout les ambiguïtés rencontrées. Ensuite, un module d'analyse grammaticale superficiel détecte les frontières des groupes nominaux en se basant sur la catégorie grammaticale des termes. Ces groupes nominaux sont décomposés, dans un premier temps en ne prenant en compte que les coordinations "et"/"ou", puis en sous-groupes nominaux en utilisant des heuristiques (environ 700). Les expressions candidates sont alors filtrées en fonction de leur fréquence. Sont éliminées les expressions apparaissant trop rarement ou trop souvent. Enfin, les éléments terminologiques ainsi sélectionnés sont validés par un expert. Celui-ci dispose d'un outils de navigation qui lui permet de voir les expressions en situation dans les textes.

Ici encore, le système est subordonné à la définition de grands volumes de connaissances explicites: une grammaire, un dictionnaire et des heuristiques.

Évaluation des résultats de l'extracteur terminologique de SPIRIT

Afin de pouvoir évaluer la performance d'une extraction automatique de la terminologie, une étude comparative entre, une extraction manuelle de KWIC faite par un expert juridique, et une extraction automatique réalisée par SPIRIT a été entreprise, afin de fournir une première évaluation grossière. Les unités documentaires prises en compte étaient les articles 229 à 310 du code civil français.

Pour l'évaluation des résultats de cette étude, deux mesures conventionnelles des systèmes de recherche d'informations ont été adoptées et adaptées, il s'agit des deux rapports : **Rappel** et **Précision** :

Rappel : ce rapport exprime la probabilité qu'un concept pertinent soit extrait. Son optimisation vise à extraire tous les concepts pertinents des unités documentaires.

$$\text{Rappel} = \frac{\text{Nombre de concepts extraits jugés pertinents}}{\text{Nombre total de concepts relevés par l' expert}}$$

Précision: Il désigne la probabilité qu'un concept extrait soit pertinent. Optimiser la précision revient à n'extraire des unités documentaires que des concepts pertinents.

$$\text{Précision} = \frac{\text{Nombre de concepts extraits jugés pertinents}}{\text{Nombre total de concepts extraits}}$$

Pour notre exemple, les résultats sont les suivants :

- Nombre total de concepts relevés par l'expert = 383

- Nombre de concepts extraits par SPIRIT et jugés pertinents = 221
- Nombre total de concepts extraits par SPIRIT = 416

Ce qui nous donne : Rappel = 221/383 = 0.58
Précision = 221/416 = 0.53

Ce taux de rappel particulièrement moyen, s'expliquerait par le fait que l'extracteur terminologique de SPIRIT ne considère que des formes nominales alors que certaines formes verbales peuvent exprimer un concept juridique pertinent. De plus, les structures syntaxiques sur lesquelles s'appuie SPIRIT pour extraire la terminologie ne tiennent pas compte des simples substantifs qui, dans le domaine juridique, constituent des concepts à part entière.

Quant à la valeur de la précision, elle s'expliquerait par le fait suivant: Généralement, pour que ces techniques d'extraction automatique de la terminologie arrivent à fournir un rapport de précision optimum, ils sont souvent utilisés sur des corpus textuels importants, ce qui permet le filtrage des expressions candidates en fonction de leur fréquence, donc, de ne retenir que des unités terminologiques pertinentes. Dans le domaine juridique, ce n'est pas uniquement la fréquence d'un concept qui détermine sa pertinence, mais c'est plutôt sa signification par rapport à la matière juridique traitée. D'autre part, il est certes raisonnable de considérer qu'un concept assez fréquent est par conséquent pertinent, toutefois, le caractère relativement spécial ainsi que la taille généralement petite des unités documentaires juridiques contrarient la véracité de cette hypothèse. Compte tenu de ces réflexions, nous avons gardé toutes les expressions fournies par l'extracteur terminologique de SPIRIT, ce choix exhaustif a par conséquent influencé le rapport de précision.

3.3.2. Construction automatique de structures classificatoires

Cette partie de l'exposé n'a pas pour but de faire une étude complète sur les types de classifications et les différents algorithmes de classifications automatiques (pour cela, on pourra consulter par exemple [Ben80] et [Vog89]). Il s'agit plutôt, de proposer quelques Méthodes inspirées de certains travaux de recherche, et de voir comment on peut les adapter afin de construire automatiquement des structures classificatoires juridiques.

DISCAN, Logiciel d'analyse de contenu et de discours [Mar 94]

Ce logiciel, conçu par Pierre MARANDA[*] permet deux types d'analyses sur un corpus textuel donné :

- Une analyse quantitative du contenu, qui après découpage du texte en unités lexicales fournit un ensemble de statistiques sur leur fréquence, leur pourcentage d'apparition, etc.

- Une analyse de discours qui fournit un graphe de concepts permettant de relever le mécanisme reliant entre elles les différentes composantes sémantiques d'un corpus textuel. C'est à partir de ce graphe qu'on parvient à construire une structure classificatoire pour les unités documentaires constituant le corpus de départ. Essayons alors de voir plus en détails le fonctionnement de DISCAN pour générer ce graphe de concepts et comment le transformer en structure classificatoire :

I) Préparation du corpus pour l'analyse :

Avant de lancer une analyse de discours, le texte à analyser doit subir quelques préparations grâce au module d'analyse de contenu. En effet, une analyse de contenu peut porter soit sur un texte brut (les articles tels qu'ils apparaissent dans le code civil) soit sur un texte filtré. Pour notre exemple, le corpus textuel sera constitué des articles réduits aux KWIC sélectionnés (soit par un expert, soit par un extracteur de terminologie). Ce choix est dû au fait que la version actuelle de DISCAN se contente de découper le texte en unités lexicales simples (simple mot délimité par des blancs ou des caractères de ponctuation). Pour contourner ce problème, les différents mots composant un KWIC donné seront reliés entre eux par un caractère spécial, ainsi, pour DISCAN l'unité lexicale sera le KWIC.

II) Construction d'un thesaurus :

[*] Professeur d'Anthropologie à l'université LAVAL, Québec, CANADA

C'est l'étape la plus délicate de la procédure. A chaque unité lexicale du corpus de départ, l'analyste associe un et un seul descripteur. Il s'agit de réduire les diversités lexicales des formes brutes (les KWIC) à des classes sémantiques (appelées aussi champs sémantiques). Ces descripteurs ne sont autres que nos fameux KWOC et KWAC.

Il est important de noter que la pertinence des résultats obtenus à la suite de la phase d'analyse de discours dépend énormément de la qualité du thesaurus. L'analyste doit donc choisir minutieusement ses descripteurs en fonction des buts qu'il souhaite atteindre.

III) "Tagging" du corpus :

Cette phase consiste à substituer les termes bruts du corpus de départ par les descripteurs correspondants. Le résultat de cette substitution est un nouveau corpus généralement appelé "Niveau secondaire" ou "Corpus normalisé".

IV) Analyse de discours :

L'analyste est maintenant en mesure de déclencher la phase d'analyse de discours. Pour ce faire, DISCAN utilise des techniques d'**analyse de Markov**. Elles consistent à calculer la probabilité de transition d'un "état" à un autre. On entend par "état", l'un des descripteur définis dans le thesaurus. Cette phase d'analyse ne porte pas sur les formes brutes du corpus de départ mais plutôt sur leur contenu sémantique (exprimé par les descripteurs du corpus normalisé).

Le résultat de cette analyse de discours est un graphe de concepts pondéré et orienté, où les sommets représentent les descripteurs (ou concepts) et les arêtes expriment leur cooccurrence[*]. Le poids associé à chaque arête représente la probabilité de succession des descripteurs dans le corpus normalisé. La position d'un descripteur par rapport à l'autre est indiquée par le sens de la flèche que constitue l'arête.

L'analyse de discours fournit également d'autres données intéressantes telle que la force sémantique d'un noeud exprimée par les deux mesures complémentaires suivantes :

- Le degré de Diffraction/Absorption: C'est le rapport arcs sortants / arcs entrants noté $\mathbf{d_+/d_-[i]}$ et interprété comme suit :
  $d_+/d_-[i] > 1$: $Noeud_i$ est un noeud diffracter (Diffraction: il émet plus qu'il ne reçoit)
  $d_+/d_-[i] < 1$: $Noeud_i$ est un noeud absorbant (Absorption: il reçoit plus qu'il n'émet)
  $d_+/d_-[i] = 1$: $Noeud_i$ est un noeud transmetteur (transmission)

- Le degré d'activité : C'est la somme du produit des arcs entrants par leurs fréquences et du produit des arcs sortants par les leurs :

$$\mathbf{(d_- * f) + (d_+ * f)}$$

Le noeud le plus actif du graphe est celui qui possède le degré d'activité le plus élevé.

V) Construction d'une structure classificatoire :

Une structure classificatoire est établie à partir d'un graphe de concepts par construction de l'arbre recouvrant correspondant. Un arbre recouvrant pour un graphe est un arbre libre joignant tous ses sommets. Cette étape qui n'est pas incluse dans DISCAN consiste, tout d'abord, à optimiser le graphe de concepts en éliminant les arêtes indésirables (les cycles, les arêtes à faibles poids, etc.), ensuite, un algorithme de génération d'arbre recouvrant est appliqué, pour enfin obtenir une structure classificatoire. Généralement, le sommet de la classification obtenue est le noeud le plus actif du graphe (ayant le degré d'activité le plus élevé).

La méthode de la proximité sémantique [Bar92]

Comme pour l'analyse de discours de DISCAN, cette méthode a également pour objectif la construction de graphes de concepts à partir d'un corpus textuel. Utilisant le logiciel SPIRIT, celle-ci procède en deux étapes :

[*] La notion de cooccurrence se traduit par la proximité spatiale entre deux descripteurs sur l'échelle linéaire que constitue le corpus normalisé

- L'extraction automatique des concepts fondamentaux grâce à l'extracteur terminologique de SPIRIT.
- La structuration des concepts extraits en graphe.

Pour cette méthode, construire un graphe de concepts consiste à déterminer les liens qui les unissent. La mesure fondamentale qui permet de détecter ces liens est basée sur la notion de cooccurrence. Pour chaque concept, on construit le "champ sémantique" correspondant. Un "champ sémantique" est défini comme étant l'ensemble des concepts qui cooccurrent avec le concept de référence dans une "fenêtre" donnée. Par "fenêtre" on désigne un ensemble de mots adjacents dans un même document. Le nombre de mots constituant une fenêtre est arbitraire (on peut le fixer à 100 mots, 200 mots, un paragraphe, etc.). La définition des liens entre différents concepts se base sur les mesures d'inclusion de leurs champs sémantiques. Considérons l'exemple de la figure 3.11.

Etant donné un concept C_1 ayant un champ sémantique CS_1 et un concept C_2 ayant un champ sémantique CS, on dit que:

$\rightarrow C_1$ est **synonyme de** C_2 **SI** $CS_1 \cap CS_2 \approx CS_1 \ \Lambda \ CS_1 \cap CS_2 \approx CS_2$ (l'intersection de CS_1 et CS_2 est presque égale à CS_1 et inversement

$\rightarrow C_1$ **chapeaute** C_2 **SI** $CS_2 \subset CS_1 \ \Lambda \ C_{cs2,cs1}$ est Important (CS_2 est inclus dans CS_1 et le complémentaire de CS_2 dans CS_1 est important)

Figure 3.11 : Quelques règles d'établissement de liens entre deux concepts

Finalement, c'est l'arbre recouvrant du graphe de concepts obtenu qui constitue la structure classificatoire, le sommet de la classification est généralement le concept ayant le champ sémantique le plus important.

3.4. Conception et codage

Un atelier logiciel supportant les deux méthodologies de construction de langages documentaires est en cours de développement. Il s'agit d'un ensemble de modules où chacun sera chargé d'accomplir l'une des différentes étapes exposées antérieurement. On peut citer à titre d'exemple :

- Module d'extraction de KWIC : il permet d'extraire dans une unité documentaire les KWIC jugés pertinents pour la construction de langages documentaires. Cette extraction de KWIC peut être faite de deux manières: soit une extraction manuelle supportée par un traitement de texte classique, soit une extraction automatique assurée par l'un des extracteurs terminologiques cités précédemment.

- Module d'affectation de KWOC : il permet d'affecter pour chaque KWIC extrait, le ou les KWOC correspondants. Ces KWOC peuvent être, soit tapés directement par l'expert, soit sélectionnés dans un thesaurus juridique.

- Module de construction des structures classificatoires : c'est l'un des plus importants modules de l'atelier, il consiste à organiser les KWIC, KWOC et KWAC sélectionnés en structures classificatoires. Pour ce faire, deux approches sont envisageables :

- Une approche manuelle qui offre à l'expert un outil adapté lui permettant de structurer les expressions sélectionnées en arborescence.
- Une approche automatique opérant en trois étapes :
    1. La génération d'un graphe de concepts soit par une analyse Markovienne (à la DISCAN) des unités documentaires soit par la méthode de la proximité sémantique.
    2. L'optimisation du graphe de concepts obtenu par élimination des arcs indésirables.
    3. La génération de l'arbre recouvrant pour ce graphe de concepts.

Le langage de programmation utilisé pour le développement de cet atelier logiciel est le C++. Un outil de conception d'interfaces graphiques appelé ZINC est également utilisé, cet outil devra garantir la portabilité de notre logiciel sous plusieurs environnements: DOS, WINDOWS, MACINTOSH et UNIX.

4. Un Atelier de Génie Documentaire Juridique (AGDJ)

Dans tout système documentaire le problème fondamental consiste à établir une communication entre un individu et un fichier. Le but en est de permettre le repérage d'une information, la sélection d'un document estimé pertinent. Un tel système repose en gros sur deux actions, le stockage et la recherche de l'information.

Si l'on essaye d'expertiser d'un point de vue purement documentaire les solutions proposées précédemment, on constate que leur intégration dans une chaîne documentaire constitue un véritable AGDJ. En effet, pour JD, l'outil d'aide à la rédaction d'abstracts (qui n'a pas été présenté dans cet article), associé à un module d'indexation, représente le maillon "Stockage" de l'information de la chaîne documentaire décrite ci-dessus. Par indexation, on entend la mise en évidence d'éléments informatifs concernant directement le contenu d'un document. Il s'agit d'une analyse du document qui tente d'en faire ressortir des indications précises quant aux idées qu'il véhicule; par exemple, les thèmes abordés, la façon dont ils sont traités, etc.

Quant à l'interface intelligente d'interrogation, elle illustre une meilleure jonction entre l'utilisateur et la base documentaire. C'est à travers cette interface que l'utilisateur procède à la "recherche" de l'information qui l'intéresse.

Indexation et recherche s'établissent au moyen de langages documentaires. Leurs vocabulaires ont pour rôle de normaliser le contenu informationnel des documents. Elles permettent de structurer tant les documents eux-mêmes, que la base documentaire et son organisation générale.

Structurer les documents revient à les organiser de façon à pouvoir les retrouver facilement et si possible individuellement au milieu de la base. Cette structuration passe éventuellement par le découpage du document en différentes zones appelées généralement "champs"; par exemple, nom de l'auteur, titre, source d'édition, etc. (ce choix est souvent adopté par les bases documentaires fondées sur la méthode de l'analyse et abstract) mais surtout par l'identification dans un document de certains éléments significatifs. Ceux-ci servent ensuite de clés pour retrouver ce document au sein d'une collection. Il s'agit là du but essentiel de l'archivage d'où l'intérêt de notre méthodologie de construction d'index.

Structurer la base ressemble plus à un effet secondaire. Cela n'est pas la raison d'être principale des langages documentaires, mais c'est une possibilité très appréciable qu'ils offrent. Les plus importants concepts informationnels, au niveau des documents, sont extraits, normalisés puis classifiés de manière à fournir des structures classificatoires des documents de la base. Ces classifications, construites grâce à la méthodologie de construction de structures classificatoires et intégrées dans une base de connaissances, servent de support pour notre interface intelligente d'interrogation.

Disposant d'un outil d'aide à la rédaction d'abstract (cas de JD), d'une interface intelligente de recherche d'information et d'outils supportant la méthodologie de construction de langages documentaires (structures classificatoires et index), on peut ainsi prétendre offrir aux documentalistes l'ossature d'un véritable Atelier de Génie Documentaire Juridique (cf. figure 4.1). Ces derniers auront toujours le soin de s'occuper du suivi et de la mise à jour de la base documentaire juridique.

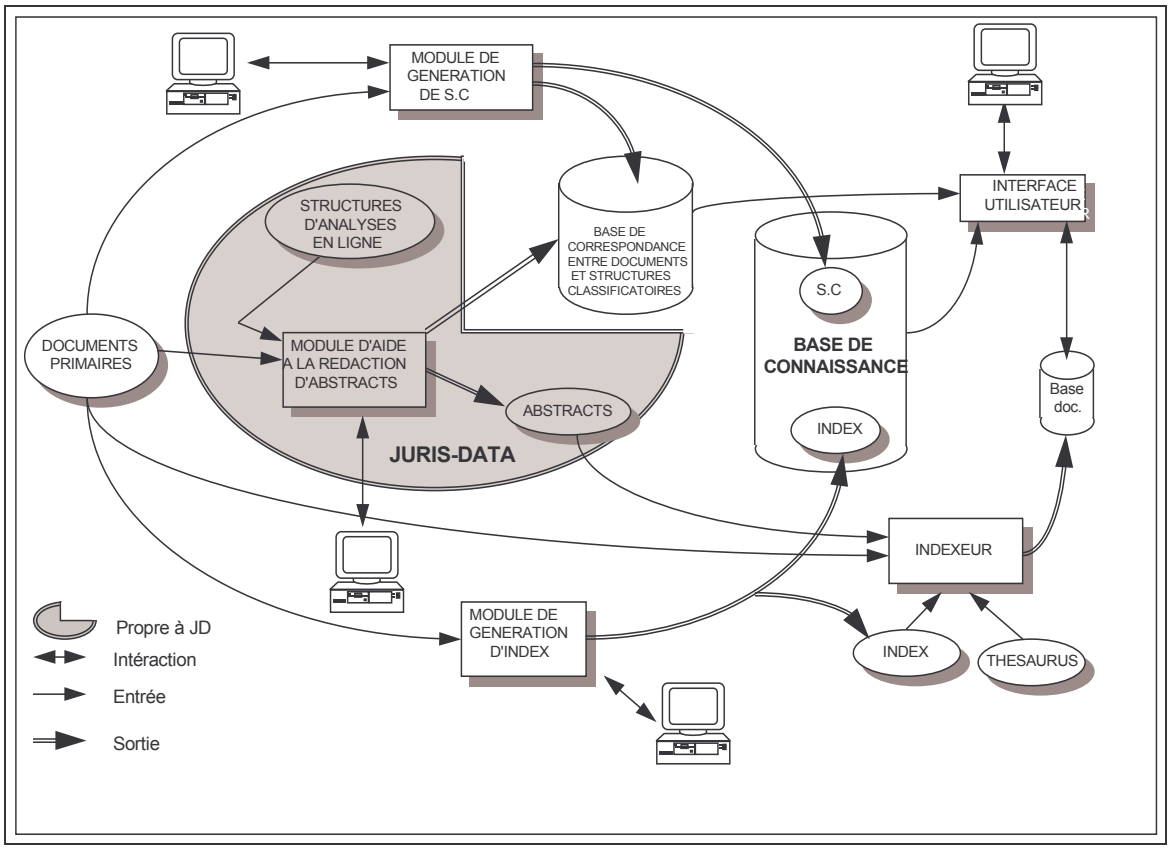

Figure 4.1 : Architecture de l'atelier de génie documentaire juridique

5. Conclusion et perspectives

Les documents des juristes, lois, jugements et arrêts, articles et commentaires, ne sont ni des documents bibliographiques, ni des documents scientifiques, ni des documents techniques ou commerciaux; en conséquence les logiciels et les réflexions de caractère documentaire existant à propos de ces derniers ne pouvaient être transposés.

Le domaine juridique étant riche en connaissances, la compétence des documentalistes s'arrêtant au seuil de la connaissance, il importe par conséquent de penser à une autre conception cognitive de la documentation juridique.

Nous avons donc développé en collaboration avec un expert juriste Mr C. Belair l'ossature d'un véritable Atelier de Génie Documentaire Juridique (AGDJ). Cet AGDJ comprend :

1) Un système d'aide à la rédaction d'abstracts pour la base documentaire juridique JD.
2) Une interface intelligente et générique pour l'aide à la recherche d'informations. Cette interface s'appuie sur une base de connaissances propre à la base documentaire juridique que l'on souhaite gérer.
3) Pour les bases documentaires juridiques non structurées, une méthodologie semi automatique de construction de langages documentaires (Structures Classificatoires et index) à partir des documents a été développée et implémentée. Structures classificatoires et index servent à la génération de la base de connaissances utilisée par l'interface de recherche d'informations.

Un tel AGDJ devra permettre d'intégrer toute la chaîne de l'information juridique depuis la production, la gestion (suivi et mise à jour) jusqu'à une diffusion plus intelligente.

Références

[Alk94] B. Alkhatib, B. Bergeon, J-L. Ermine, C-M. Falinower, M. Monsion : *Génie logiciel et Génie cognitif pour l'élaboration d'une base de connaissances en automatique*, 9ième congrès Reconnaissance des formes et Intelligence Artificielle, RFIA'94, Paris 11-14 Janvier 1994, Vol 2, pp. 734 - 738, Paris, 1994

[And83a] Andreewsky A., Binquet, Debili F., Fluhr C., Ponderoux : *L'interrogation en langage naturel dans le système SPIRIT*, Journées Internationales de l'Informatique et de l'Automatisme, pp. 322-332, 1983.

[And 83b] Andreewsky A., Debili F., Fluhr C. : *Apprentissage - syntaxe - sémantique lexicale*, Revue du Palais de la découverte, Vol. 9, N°83, décembre, 1983

[Bar92] Barakat Barbieri B. :Vers une c*onstruction automatique de graphes de concepts*, Thèse de Doctorat de l'École Centrale, 1992

[Ben80] Benzécri J.P : *L'analyse des données, Tome 1 : la taxinomie* - Éditions DUNOD, 1980

[Bou91] Bourcier D.: *Méthodes pour une approche cognitive du droit*, Les sciences cognitives en débat, B. Vergnaud Ed, CNRS Éditions 1991

[Bou92] Bourigault D. : *Lexter, vers un outils linguistique d'aide à l'acquisition des connaissances*, 3èmes journées d'Acquisition des connaissances du PRC-IA, Dourdan, Avril, 1992

[Bru94] Brunet E., Ermine J.-L. : *Problématique de la Gestion des Connaissances des Organisations*, Ingénierie des systèmes d'information, Vol. 2, n° 3, pp. 263-291, AFCET/Hermès, 1994

[Dav] David S., Plante P.: *De la nécessité d'une approche morphosyntaxique en analyse de texte*, 25 pages, Rapport interne UOAM, Québec.

[Erm89] Ermine J.-L. : *Systèmes experts, théorie et pratique*, Collection Tec et Doc, Lavoisier Ed, Paris, 1989

[Erm93] Ermine J.-L. : *Génie Logiciel et Génie Cognitif pour les systèmes à base de connaissances*, Collection Tec et Doc, Lavoisier Ed, Paris, 1993

[Gre66] Greimas A-J : *Sémantique structurale*, Larousse, Paris, 1966 (repris chez P.U.F. Paris, 1986)

[Hic 89] Hickman F.R., Killin J., Land L., Mulhall T., Porter D., and Taylor R.M.: *Analysis for knowledge based systems, a practical guide to the KADS methodology*, Ellis Horwood books in information technology, 1989

[IGL89] IGL Technology: *SADT un langage pour communiquer*, 1989 - Éditions EYROLLES

[Mar 94] Maranda P., Nze-Nguema F-P. : *L'unité dans la diversité culturelle,* Les presses de l'Université Laval, Sainte-Foy, Québec, 1994

[Mic 83] R.S. Michalski, J.G. Carbonell, T.M Mitchell. : *A theory and methodology of inductive learning*, Eds, Machine learning: an artificial intelligence approach, Tyoga, pp. 83-129, 1983

[Sal 72] Salton G.: *A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART)*, Journ. of the Americ, Soc. for Inf. Sci. Vol. 23, N° 2, March-April, pp. 75-84, 1972

[Sca 90] Scapin D., Pierret-Golbreich C.: *Towards a method for task description : MAD*, Work with display units 89, L. Berlinguet, D. Berthelett Eds, Elsevier Science, North Holland Publishers, 1990

[Ska 89] Skalak D.B. : *Taking advantage of models for legal classification*, Second International Conference on Artificial Intelligence and Law, pp. 234-241, ACM New York, 1989

[SPI 92] *SPIRIT of SYSTEX - the linguistic skill*, Rapport interne SYSTEX, Bâtiment Appolo, Espace Technologique, 91195, Saint-Aubin cedex, France, 2 pages, 1992.

[Vog 88] Vogel C. : *Génie cognitif*, - Éditions MASSON, Coll. Sciences cognitives, 1988.